DOCUMENT RESUME

ED 338 262 IR 053 811

AUTHOR	Fox, Edward A.; Wilson, Linda
TITLE	Comparison of Advanced Retrieval Approaches for Online Catalog Access. Final Report.
INSTITUTION	Virginia Polytechnic Inst. and State Univ., Blacksburg. Univ. Libraries.
SPONS AGENCY	Council on Library Resources, Inc., Washington, D.C.;
PUB DATE	91
NOTE	21p.
PUB TYPE	Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Libraries; Analysis of Variance; Higher Education; Hypothesis Testing; *Information Retrieval; *Online Catalogs; *Online Searching; Relevance (Information Retrieval); *Search Strategies

ABSTRACT

This report presents the results of an experimental study which compared four different advanced retrieval methods for an online catalog to determine which are more efficient, effective, and usable. Conducted with 216 student volunteers from Virginia Polytechnic Institute and State University (Virginia Tech), the study required access to data for a large number of catalog records, a software system that could be used to search through that data, a computer system that could support a number of simultaneous users searching against the catalog database, participants who would try out the system on the records, and questions asked by students in the library. Eighteen of the questions collected using a specially designed form were selected to be searched by the participants and later by expert librarians. The students' responses to an online questionnaire and indications of timing and effectiveness from the log of the search session were used to compare the four methods. Analyses of these variables indicated that the study results provide support for the research hypothesis, i.e., that the vector with feedback method would be the most effective retrieval method, followed by the vector method, and then by the extended Boolean and Boolean methods. This report includes detailed information on the funding received; descriptions of the research design, preparing for and conducting the experiment, an overview of the results, detailed results from the questionnaire analysis, and a brief statement on further and future work. Six appendixes contain a related article, "Advanced Retrieval Methods for Online Catalogs" (Edward A. Fox); the form used for collecting queries; a description of the protocol for the Retrieval Experiment--Virginia Tech OnLine Catalog (REVTOLC); and the REVTOLC consent form, explanations and detailed instructions for the subjects, and a questionnaire which includes demographic questions. (MAB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED338262

Final Report: Comparison of Advanced Retrieval Approaches for Online Catalog Access

Edward A. Fox
Dept. of Computer Science and Computing Center

Linda Wilson
Reference Dept., Newman Library
Virginia Polytechnic Institute and State University
(VPI&SU or Virginia Tech)
Blacksburg, VA 24061

September 13, 1991

120538?

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Linda Wilson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

1 Grant Details

Grant Number: CLR 4048-C

Grant Period: 1/1/89 – 6/30/90

Grant Amount: $3000

Principal Investigators: Edward A. Fox and Linda Wilson

Graduate Research Assistant: Amjad M. Daoud

Secretary: Karen Davis

Related Project: Advanced Retrieval Methods for Online Catalogs, funded by Online Computer Library Center (OCLC) Inc., 7/1/89 – 6/30/90, $10,000, principal investigator E. A. Fox

Related Report: p. 32-34 of "Annual Review of OCLC Research," July 1989 – June 1990, OCLC, Dublin, Ohio (copy attached to this report in Appendix A)

2 Support from CLR and other Sources

An accounting of the expenditures has been prepared and submitted by our Office of Sponsored Programs. Funds were provided by CLR and by OCLC (see above).

Some of the funds were used to pay subjects who participated in our experiment (220 people each received $5 for a total of $1100). Other funds were used to pay for time spent by the secretary, Karen Davis, who helped with the experiment and with subsequent data analysis. OCLC funds were used to support the efforts of the graduate research assistant, Amjad M. Daoud. Mr. Daoud, progressing toward his Ph.D., was later paid by the Virginia Tech Computing Center while he worked on other phases of this effort. In addition, he contributed a great deal of his own time without pay. Significant costs in terms of computer time (only $300 was charged while many thousands of dollars worth of computing were used) and materials and supplies (tapes, paper, copies, etc.) were also incurred and covered by the University.

The two principal investigators were employed by Virginia Tech during this investigation and received no extra funds for this project, but instead worked on this in addition to their other areas of responsibility. Volunteer assistance was also provided by several on the staff in Newman Library, who supplied expert judgments regarding relevance of various retrieved records from the online catalog in relation to the queries being considered.

Altogether, 3-4 person years of work have been devoted to this project, and the real cost of the effort to-date would probably exceed $200,000.

3 Experiment

Most of the work relating to this grant involved an experimental study, that was designed, prototyped, carried out, and analyzed.

3.1 Design

This project considered searching online public access catalogs using four different retrieval methods:

1. Vector (where a list of words is entered and the computer returns a ranked list of close matches)
2. Vector with feedback (where after vector retrieval takes place, the user's indication of which documents are relevant to the query is considered, and the system can automatically construct a "feedback" query that will also yield a ranked list of close matches)
3. Boolean
4. Extended Boolean (actually, the "p-norm" implementation of this general approach, where Boolean queries are formed but the computer "softens" the interpretation of the "and" and "or" operators to allow close matches and ranked output)

To effect a realistic comparison, the following ingredients were necessary:

- access to data for a large number of catalog records — in our case about 500,000 records from the Newman Library VTLS system were included.
- a software system that could be used to search through that data — the SMART system (originally developed at Cornell) was modified for this purpose, with a great deal of interface software added to support library record manipulation.
- a computer system that could support a number of simultaneous users searching against the catalog database — a Sequent Symmetry in the Department of Computer Science was used, with its 10 processors, along with related network connections, terminals, and other networked computers.
- participants who would try out the system on the records — over 220 students at Virginia Tech were recruited
- questions asked by students in the library — many were collected using a specially designed form (see Appendix B). From these, 18 were selected and searched by the participants and also later by expert librarians. Two typical questions expressed as interest statements are numbers 5 and 7:
  - 5. I need to know more about endangered species and fisheries ecology.
  - 7. I need information about remote sensing.

Our hypothesis was that the vector with feedback method would be best, followed by the vector method, and then by the extended Boolean and Boolean methods. To determine if this was true, we sought to isolate the effect of retrieval method from the effects of other variables. The other variables considered were:

- question — Would results vary based on the question considered?
- participant — Would results vary between the participants?

- order of searching — Since we decided that each partipant should search for records in response to two queries for each of two retrieval methods (i.e., 4 queries per participant), would the order of the two methods influence the results?

Our design called for 12 different orders (4 methods for the first two queries and 3 other methods for the remaining queries) and 18 queries, so we had 216 (12*18) participants.

To compare the four methods, some measures had to be identified. Two sources of data were available: the replies of participants to (online) questionnaires, and the log of the search session. This latter source could be used to provide both timing and effectiveness indications.

3.2 Preparing for the Experiment

An earlier version of this study was carried out in 1987, with 52 partipants, 300,000 records, and a relatively difficult to use interface. Based on lessons learned, a great deal of development and testing work was carried out, especially to improve the interface. Of particular importance was the fact that all four search methods used a very similar forms-style screen layout, so that the effect of the interface on retrieval results and preferences would be reduced greatly.

In addition, to handle the larger number of participants, the SMART system was ported from our VAX 11/785 uni-processor system to a 10 processor Sequent, allowing up to 10 partipants to search at the same time without degrading performance. Performance tuning work was undertaken to speed up the search, so that each of the four methods being tested would have similar performance.

A larger collection (500,000 instead of 300,000) of catalog records was loaded onto the computer. Also, 18 questions were selected from a larger set of questions collected over a several month period, at a few locations in Newman Library.

Online help and written training materials were prepared and tested. A pilot study was done with four students (with $1100 allocated, and $5 for each participant, there was money for 216 participants and 4 pilot testers). By the summer of 1990, everything was ready to carry out the main study.

3.3 Conducting the Experiment

In the spring of 1990, permission was obtained from the University board overseeing research involving human subjects, based on our explanation of the protocol, which appears in Appendix C. During the summer of 1990, all 216 student participants were recruited and involved in the experiment. Sessions, in which 1-8 people appeared, were scheduled for 2 hour periods, though participants generally completed their work within an hour.

The graduate research assistant, secretary, and principal investigators were each present at varying number of sessions. Since interaction with the session "proctor" was minimal, and since the graduate assistant was present in the vast majority of the sessions, it was not deemed necessary to control for the effects of the proctor.

During each session, participants were asked to read and sign a consent form, illustrated in Appendix D. They also read an explanation of the experiment, similar to that shown in Appendix E (with the "blanks" filled in for their particular session). As they carried out the instructions given there, they completed a variety of online questionnaires, with the questions summarized in Appendix F.

4 Overview of Results

All needed data was collected. However, it quickly became clear to the graduate research assistant that data analysis would be an enormous undertaking. Hence, with the aid of the secretary (who had completed an M.S. in statistics), analysis of the questionnaire data was undertaken first. However, the secretary took up other employment at the end of the summer of 1990. Also, the graduate assistant was forced (because of economic effects of the Gulf War) to take a leave of absence from his graduate studies, beginning very early in 1991.

While a detailed writeup still needs to be completed, much of the questionnaire analysis effort is complete. In most cases the vector and vector with feedback cases give similar results, grouping together. Similarly, the Boolean and extended Boolean cases also gave similar results and could be grouped together, though it was almost always the case that the Boolean case was rated as poorest. Participants generally preferred the vector group to the Boolean group. If a "best" method had to be picked based on the questionnaires, it would be vector with feedback; the "worst" would be Boolean.

Data regarding timing of events during the searches has been processed and some of the key figures have been loaded into a spreadsheet format. However, the files are so large that transfer to a relational database system is planned, so that further analysis can be carried out.

In recent months, the evaluation package in SMART that allows computation of recall, precision, and other effectiveness measures has been made to work at Virginia Tech on other data sets. Given the large quantity and nature of the data available, a good deal of thought, computer processing, analysis and writing is needed to determine which methods were most effective in helping users find relevant items.

Principal investigator Fox was invited to the American Library Association 1991 Annual Conf., held in Atlanta, GA, to give a talk on "Distributed Expert Based Retrieval and OPACs" on July 2, 1991. Some preliminary comments on this project were made, covering some of the observations mentioned above.

5 Detailed Results from Questionnaire Analysis

For each of the questions in Appendix F, results were compared as a function of the four retrieval methods:

1. Vector
2. Vector with automatic feedback
3. Boolean
4. P-norm

An analysis was done using one-way ANOVAs with the SAS package. No significant differences were found regarding questions 1-12, 28, 30, 31. Other questions showed significant differences in responses, depending on which method was used. For those questions, the means are shown in Table 1.

The tests indicated many significant differences. For question:

13 — regarding computer time delay in response, there was a significant difference between methods 1&2 (which people found shorter) and methods 3&4 (i.e., 1 vs. 3, 2 vs. 3, 1 vs. 4, 2 vs. 4).

Table 1: Means for Questions for 4 Methods					
Question Number	Means for 4 Retrieval Methods				Value Range and Description
	1 (Vector)	2 (Fdbk)	3 (Boolean)	4 (P-Norm)	
13	2.87	2.99	2.51	2.54	1:long - 4:short
14	2.18	2.12	2.69	2.33	1:more - 4:nothing
15	2.68	2.83	2.47	2.86	1:0-20 - 5:80-100%
16	1.73	1.70	2.27	1.89	1:easy - 4:difficult
17	2.22	2.29	1.71	2.08	1:too much - 3:less
18	2.65	2.36	2.94	2.55	1:st.agree - 5:st.disag.
19	2.25	2.14	3.16	2.62	1:very eff. - 5:not eff.
20	2.14	2.31	3.11	2.51	1:very eff. - 5:not eff.
21	1.63	1.55	1.82	1.74	1:st.agree - 5:st.disag.
22	2.15	2.09	2.54	2.35	1:st.agree - 5:st.disag.
23	1.10	1.09	1.32	1.20	1:yes - 2:no
24	1.89	1.84	2.69	2.45	1:st.agree - 5:st.disag.
25	1.92	1.84	2.86	2.40	1:st.agree - 5:st.disag.
26	2.57	2.28	3.31	2.99	1:st.agree - 5:st.disag.
27	2.22	2.08	2.84	2.32	1:st.agree - 5:st.disag.
29	1.87	2.01	2.65	2.27	1:st.agree - 5:st.disag.

14 — regarding the number of useful items found, there was a significant difference between methods 1,2&4 (which people found yielded more) and method 3.

15 — regarding the percentage of the documents found that were relevant, there was a significant difference between method 4 (yielding a higher percentage) and method 3.

16 — regarding ease of search, there was a significant difference between methods 1,2&4 (which people found easier) and method 3.

17 — regarding time for a search, there was a significant difference between methods 1,2&4 (which people found took less) and method 3 — and also between 4 (less) and 3.

18 — regarding if the results of the search were what was expected, there was a significant difference between methods 2&4 (agree) and method 3.

19 — regarding effectiveness of helping casual searching, there was a significant difference between methods 1,2&4 (which people found effective) and method 3 — and also between 4 (effective) and 3.

20 — regarding effectiveness of helping comprehensive searching, the means are in order (from most effective to least): 1, 2, 4, 3 with significant differences 1 vs. 3&4, 2 vs. 3, 4 vs. 3.

21 — regarding if it was helpful to manually edit earlier queries, there was a significant difference between methods 2 (agree) and 3.

22 — regarding if the display of details was clear and adequate, there was a significant difference between methods 1&2 (agree) and method 3.

23 — regarding if the information was easy to understand and adequate, there was a significant difference between methods 1&2 (yes) and method 3.

24 — regarding if the method was easy to learn, there was a significant difference between methods 1&2 (agree) and methods 3&4 (i.e., 1 vs. 3, 2 vs. 3, 1 vs. 4, 2 vs. 4).

25 — regarding if the method was easy to use, there was a significant difference between methods 1&2 (agree) and methods 3&4 (i.e., 1 vs. 3, 2 vs. 3, 1 vs. 4, 2 vs. 4) — also between methods 4 (agree) and 3.

26 — regarding if the method was easier than VTLS (the locally used library catalog system), there was a significant difference between methods 1&2 (agree) and method 3 — also between methods 2 (agree) and 4.

27 — regarding if the method did a good job, there was a significant difference between methods 1&2 (agree) and method 3 — also between methods 4 (agree) and 3.

29 — regarding if the editor was easy to use, there was a significant difference between methods 1&2 (which people found faster) and methods 3&4 (i.e., 1 vs. 3, 2 vs. 3, 1 vs. 4, 2 vs. 4).

Clearly, method 3 (Boolean) is not well liked, and method 2 (vector with relevance feedback) is generally preferred.

6 Further and Future Work

Since the graduate research assistant in this project will be reporting on the experiment in his Ph.D. dissertation, it is clear that there will be further work on analysis and description of the effort. Ultimately, at least one journal article will be prepared to disseminate findings.

Based in part on the preliminary findings from this study, a related development effort is now underway at Virginia Tech. We have completed a design for the MARIAN (Multiple Access Retrieval of library Information with ANnotation) system, for searching online library catalogs, using many of the ideas explored in this project. Programming is proceeding. It is hoped that by Spring 1992, the entire Virginia Tech file of catalog data will be loaded onto two NeXT computers, that will act as search servers, supporting at least vector and vector with feedback searching by many simultaneous users.

As an interim measure, another graduate student, Nasser Ghazi, undertook development of a prototype of one interface for the MARIAN system, using Motif and the X Window System. He also developed client-server software to work with the SMART system, so that multiple search sessions can be carried out simultaneously. This effort is reported in his M.S. report, a copy of which will be provided to CLR. It benefited from the REVTOLC study, which he also discusses briefly in his report.

The REVTOLC study has provided a great deal of insight to the investigators, regarding how to process catalog data and how university students react to advanced retrieval methods. Future articles will explore these issues in more detail.

APPENDIX

Table of Contents

- A. Article in "Annual Review of OCLC Research"
- B. Form for Collecting Queries
- C. Protocol
- D. Consent Form
- E. Explanation
- F. Questions

Advanced Retrieval Methods for Online Catalogs

Principal Investigator: Edward A. Fox, Associate Professor, Department of Computer Science, Virginia Polytechnic Institute and State University

Abstract

This project has dealt with an important problem facing libraries: How can we improve search methods for online catalogs so they are both simpler to use and more effective? At the same time we have addressed a problem facing the information retrieval community: Can advanced retrieval methods be successfully applied to large collections? Our experiment involves 216 students as subjects, four retrieval methods, 18 questions collected from library users, and roughly half a million catalog records from the VPI&SU library catalog. To handle large collections, we developed fast methods for file inversion and perfect hashing techniques suitable for millions of keywords. We have greatly extended the SMART retrieval system, added a forms-type interface, and used it for Boolean, extended Boolean, vector, and vector with probabilistic feedback searching. Preliminary observations indicate that students prefer the vector schemes.

Many libraries now use online catalogs to help users locate books and other relevant materials. Search techniques generally include one or more of the following: (1) traversing menus, (2) exact matching of titles or title prefixes, and (3) submitting Boolean queries. Yet, while it is well known that going from an information need or question to a suitable Boolean query is very difficult, the other two methods have limited applicability and power. How can catalog searching be made easier and more effective?

We investigated the applicability of advanced retrieval methods with a large online library catalog collection. Our first task was to ensure that the necessary supporting technology was developed. We began by studying techniques for handling the large number of keys (names, terms, and keywords) commonly found.

Minimal Perfect Hashing

In library catalogs, the number of keys is on the order of hundreds of thousands or millions, so secondary storage devices (magnetic disks or CD-ROMs) are frequently used. We explored the use of perfect hash functions, which can locate any allowable key in a single disk access to improve system efficiency. At the same time, to reduce storage overhead, we sought minimal perfect hash functions, which guarantee 100% loading of hash tables.

In earlier work we improved upon published algorithms that could find minimal perfect hash functions in O(n**4) time, with a new method that only required O(n**3) time (Fox, Chen, Heath, and Datta 1989). Thus, to find a minimal perfect hash function for 10 keys, our method would require on the order of 1,000 time units instead of 10,000 time units. Since this is still not suitable for library collections, we improved our method to require only O(n log n) time (Fox, Heath, and Chen 1989).

In connection with this project, we extended our work even further, as part of the doctoral research of Amjad Daoud, so minimal perfect hash functions can be found with high probability in O(n) time. Furthermore, we developed a method for finding order-preserving minimal perfect hash functions (Fox, Chen, Daoud, and Heath n.d.), where any a priori ordering of the keys is maintained. For static collections, this method should compete favorably with B-trees, especially now that we can find order preserving minimal perfect hash functions with high probability in O(n) time.

OCLC provided a very large set of names in machine-readable form; thus, our experiments involved tests including millions of keys. The techniques developed should be readily applicable to CD-ROM-based catalog systems, and of use in a wide variety of other situations where data changes slowly and efficient access is required.

Other Supporting Technology

To demonstrate how advanced retrieval methods could be used with large catalogs, we selected the SMART retrieval system, developed at Cornell University. Since the version we used had only been applied to collections with approximately 15,000 documents, numerous changes were required. An earlier pilot study had demonstrated the feasibility of our experiment using some 300,000 records and a crude interface. Some 52 subjects participated, and a group of students helped during the query construction process, but analysis was limited to comparing the responses to online questionnaires (Fox 1987). Considerable improvement was needed to obtain definitive results, thus leading to this current project.

First, we developed a disk-based method to build inverted files that made near optimal use of available primary memory, and required O(n) time given a collection of compressed document vectors with total length n. Second, we ported SMART to run on a ten-processor Sequent computer under Dynix, a version of the UNIX operating system. In this regard we explored various approaches to parallel computation for retrieval. Third, we improved the search, retrieval, and ranking algorithms to be considerably faster for the four retrieval methods being investigated. Fourth, we developed a new form-based interface to SMART for VT100-type terminals. The interface was relatively consistent across the four retrieval methods. Fifth, we developed logging and tutorial software so that users involved in our experiment would have a gentle introduction to searching. The software provided us with information through online questionnaires and logs of keystrokes entered during the search sessions.

Experiment

The experiment had a simple design. We compared four retrieval methods:

1. Boolean
2. Extended Boolean (i.e., pnorm)

5. Vector
4. Vector with probabilistic feedback

Each subject used two methods, and we varied the order in which the searching methods were introduced to users, yielding 12 test groups (4 × 3).

Each subject searched for matches for four different questions, randomly selected from a set of 18, which in turn were chosen from a number of forms filled in by library users in fall 1989. With 18 questions and 12 groups, we decided to have 216 subjects (18 × 12). Given that each user worked with two methods and four questions (two per method), we expect to have sufficient statistical data to compare the four methods.

Besides handling the order of searching with the two methods, we attempted to ensure the generality of our results by roughly balancing the subjects with respect to gender, academic level (from freshman to graduate student), and college (randomly selecting courses offered during the spring and first summer sessions, 1990).

Data Collection

Data collection includes replies to initial and final questionnaires, as well as questionnaires relating to each of the four searches and each of the two methods. After each query is processed, users look at a detailed screen for any retrieved document and indicate whether the document is relevant: "Y" for yes, "N" for no, or "D" for don't know. We will analyze recall/precision performance after VPI&SU librarians complete a large number of searches for each of the 18 questions.

Two questions used in the test are (1) I need to know more about endangered species and fisheries ecology, and (2) I need to know more about ceramic matrix composites.

One possible Boolean search formulation for the first question appears in figure 2-1. The form-based entry screen simplifies the task for novices, and yet gives a great deal of flexibility. Terms on each line are ANDed, but lines for each "field" can be either ANDed or ORed, and the various fields also can be either ANDed or ORed (see explanation about PF2 and PF3 in upper right of the screen). The results of this Boolean search appear in figure 2-2; the first of two detailed screens for the fourth entry appears in figure 2-3.

An example of the form-based entry for vector queries appears in figure 2-4. In this scheme, each term entered is assigned a weight, and pattern matching between the query and the documents leads to selection and ranking for the ten best matches. The results are shown in figure 2-5, captured after relevance judgments were added by the author. Users may elect to have the computer system automatically construct a new query based on relevance feedback; the results are shown in figure 2-6.

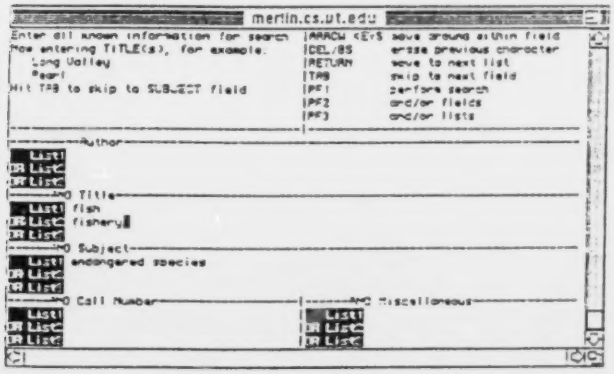

Fig. 2-1 Boolean Query

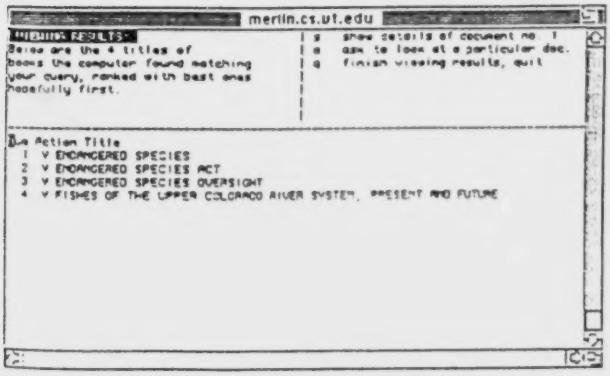
merlin.cs.ut.edu
MERLIN RESULTS:
Below are the 4 titles of
books the computer found matching
your query, ranked with best ones
hopefully first.
No Action Title
1 Y ENDANGERED SPECIES
2 Y ENDANGERED SPECIES ACT
3 Y ENDANGERED SPECIES OVERSIGHT
4 Y FISHES OF THE UPPER COLORADO RIVER SYSTEM, PRESENT AND FUTURE

Fig. 2-2 Boolean Query Results and Relevance Judgments

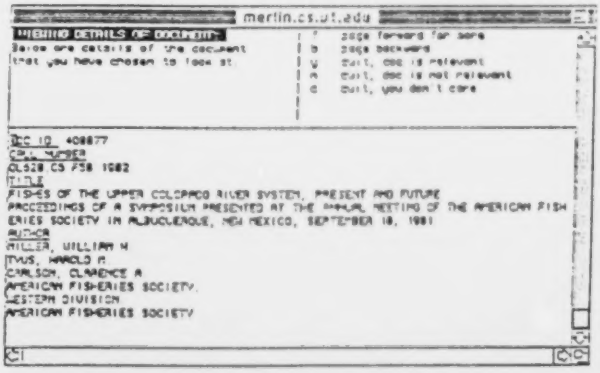
merlin.cs.ut.edu
MERLIN DETAILS OF DOCUMENT
Below are details of the document
that you have chosen to look at.
f page forward for more
b page backward
y quit, doc is relevant
n quit, doc is not relevant
d quit, you don't care
Doc ID 408877
Call Number
QL638 C6 F58 1982
Title
FISHES OF THE UPPER COLORADO RIVER SYSTEM, PRESENT AND FUTURE
PROCEEDINGS OF A SYMPOSIUM PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN FISH
ERIES SOCIETY IN ALBUQUERQUE, NEW MEXICO, SEPTEMBER 18, 1981
Author
MILLER, WILLIAM H
TYUS, HAROLD M
CARLSON, CLARENCE A
AMERICAN FISHERIES SOCIETY.
WESTERN DIVISION
AMERICAN FISHERIES SOCIETY

Fig. 2-3 First Page for Selection 4 from Boolean Query

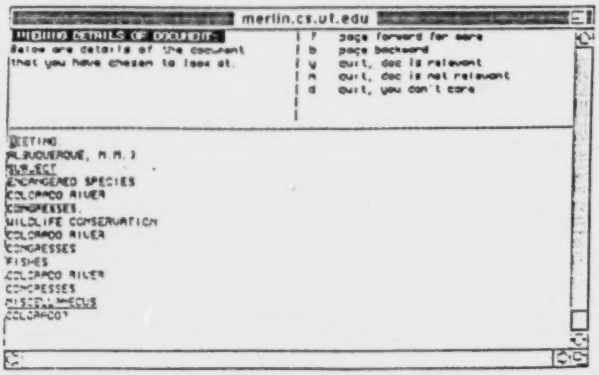

Fig. 2-4 Vector Query

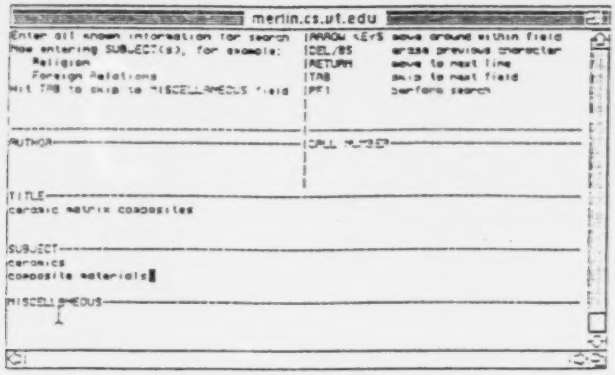

Fig. 2-5 Vector Query Results and Relevance Judgments

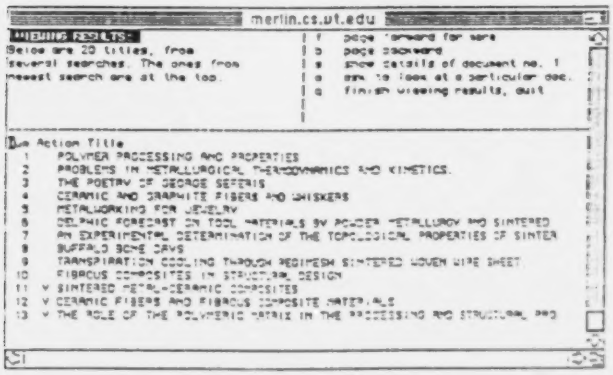
merlin.cs.ut.edu
SEARCH RESULTS:
Below are 20 titles, from several searches. The ones from newest search are at the top.
f page forward for more
b page backward
s show details of document no. 1
o ask to look at a particular doc.
q finish viewing results, quit
No Action Title
1 POLYMER PROCESSING AND PROPERTIES
2 PROBLEMS IN METALLURGICAL THERMODYNAMICS AND KINETICS.
3 THE POETRY OF GEORGE SEFERIS
4 CERAMIC AND GRAPHITE FIBERS AND WHISKERS
5 METALWORKING FOR JEWELRY
6 DELPHIC FORECAST ON TOOL MATERIALS BY POWDER METALLURGY AND SINTERED
7 AN EXPERIMENTAL DETERMINATION OF THE TOPOLOGICAL PROPERTIES OF SINTER
8 BUFFALO BONE CARVE
9 TRANSPIRATION COOLING THROUGH REGIMESH SINTERED WOVEN WIRE SHEET
10 FIBROUS COMPOSITES IN STRUCTURAL DESIGN
11 v SINTERED METAL-CERAMIC COMPOSITES
12 v CERAMIC FIBERS AND FIBROUS COMPOSITE MATERIALS
13 v THE ROLE OF THE POLYMERIC MATRIX IN THE PROCESSING AND STRUCTURAL PRO

Fig. 2-6 Vector Query Results (starting with number 11) and Feedback Results (entries 1–10)

As of June 1990, roughly half of the 216 subjects have completed their search session. Preliminary observations indicate that the vector and vector with feedback methods are easier for most users. All methods can be learned fairly quickly, in part due to the simplicity of the interface and the guidance given in three pages of written instructions and a number of tutorial screens. Most subjects have managed to complete their work with the four questions, as well as the several questionnaires, in about an hour.

Further Research

Once the experiment is complete, data will be extracted and analyzed. While the primary comparison will be of the four retrieval methods, many correlations will be studied. Effects of major, gender, academic level, computer training, experience with retrieval systems, experience with online catalogs, typing skill, degree of patience, and other attributes will be investigated.

We will continue to improve the efficiency of the hashing methods, speed up the search algorithms, explore opportunities for parallel processing, and develop a more flexible interface.

References

Fox, E. A. 1987. Testing the applicability of intelligent methods for information retrieval. *Information Services and Use* 7:119–38.

Fox, E. A., Q. F. Chen, A. M. Daoud, and L. S. Heath. n.d. Order preserving minimal perfect hash functions and information retrieval. In *Proceedings SIGIR 90, Thirteenth international conference on research and development in information retrieval, Brussels, Belgium, September 5–7, 1990.* Forthcoming.

Fox, E. A., Q. F. Chen, L. Heath, and S. Datta. 1989. A more cost effective algorithm for finding perfect hash functions. In *Proceedings ACM 1989 Computer Science Conference, Louisville, Ky., February 21–23, 1989,* 114–22.

Fox, E. A., L. Heath, and Q. F. Chen. 1989. *An O(n log n) algorithm for finding minimal perfect hash functions.* Technical Report TR-89-10. Blacksburg, Va.: Virginia Polytechnic Institute and State University, Department of Computer Science.

This study was supported by Council on Library Resources Grant CLR 4048-C, "Comparison of Advanced Retrieval Approaches for Online Catalog Access," with the assistance of co-principal investigator Linda Wilson (Reference Department, Newman Library, VPI&SU), Amjad M. Daoud, graduate research assistant, and Karen Davis, secretary.

Back-of-the-Book Index Study

Principal Investigator: Elizabeth D. Liddy, Assistant Professor, School of Information Studies, Syracuse University, New York

Abstract

Although we are familiar with back-of-the-book (BOB) indexes, surprisingly little is known about how they are produced, what they are likely to contain, and what makes one index better than another. In particular, benchmark data about the nature and format of BOB indexes are needed before the utility and feasibility of using indexes for information retrieval can be evaluated. This study, which is ongoing, examined BOB indexes of 424 recent books in a broad range of subject areas. It seeks to (1) describe characteristics and features of indexes, (2) evaluate the quality of indexes, and (3) describe current practices and policies related to the production of BOB indexes.

Currently, back-of-the-book (BOB) indexes are used to provide manual access to the contents of individual books. Current research, however, suggests the possibility of using BOB indexes to augment online catalog records, thus providing improved subject access to books (Kinnucan 1987). But before online system designers can use the detailed information in BOB indexes effectively and efficiently, they need to know more about their characteristics, arrangement, and quality. Further, it would be useful to know more about how indexes are currently constructed. The only detailed BOB index study (Gratch, Settel, and Atherton, 1978) is more than ten years old and it evaluated books with much earlier publication dates. The goal of the

B Form for Collecting Queries

PLEASE HELP YOUR LIBRARY

THE LIBRARY IS COLLECTING INFORMATION ABOUT HOW USERS SEARCH THE VTLS ONLINE CATALOG. WE WILL APPRECIATE YOUR COMPLETING THIS FORM WHILE YOU ARE SEARCHING VTLS. IMMEDIATELY AFTER, PLEASE LEAVE IT IN THE BOX LABELLED "LIBRARY SURVEYS". THANK YOU!!

Date:_____ Time:_____ Sex: M__ F__ Age: ____ Major: _____

Academic Level: F__ S __ J __ S __ Grad __

How much experience have you had using computers:
little or none __ some __ a lot __ use computers daily __

How often do you use VTLS:
daily __ several times/week __ several times/month __ several times/semester __
rarely __

Why are you using the library today (In this and all subsequent questions consider only your most important reason) :
leisure __ class assignment __ thesis __ other __
If assigment, please give:
College: _____ Department: _____
Course #: _____

How familiar are you with the Library of Congress Subject Headings:
never heard of them __ have heard of them but don't use them __
use them regularly for subject searches __ use them regularly and understand them well __

Describe your specific information need: For example, "I need information about the possible causes of homelessness in contemporary society in the United States" or "How are black holes formed?" or "I need to check out a copy of The Grapes of Wrath by John Steinbeck" or "I need a book on COBOL programming."

How many books do you intend to check out on your topic:
one __ several __ about 5 __ everything I can find __

What do you know about your subject that might help you find information on VTLS (Check all that apply):

C Protocol — REVTOLC (Retrieval Experiment - Virginia Tech OnLine Catalog)

The purpose of this experiment is to lead to more effective retrieval methods, especially for improving access to online public access catalogs. The objective of this study is to compare four methods (Boolean, extended Boolean, vector, and vector with feedback) to determine which are better, regarding efficiency, effectiveness, and usability. The actual experiment will involve roughly a half million records provided by staff of the University Library in 1986, that will be searchable using computers in the Department of Computer Science, and computer terminals in public laboratories (e.g., McBryde 116) and other locations.

Participants will number around 220, so that adequate statistical power will allow us to identify small differences between methods. Since there are 4 methods considered, and since it will be necessary for each subject to work with 2 methods, there are 4*(4-1) = 12 combinations of retrieval method pairs; each participant will be randomly assigned to one of these.

All of the participants will be introduced to the use of the system and will receive explanations regarding the process of the experiment at the beginning of the session. They will receive and be asked to sign a consent form which explains the process and informs them that they are free to quit the experiment at any time.

Each participant that completes the experiment will be paid $5 for their assistance, which should require roughly 60-90 minutes of time. A participant will read a short booklet of instructions, and then do all further work with the computer, which will log all key strokes, especially responses, adding timestamps. The steps with the computer are as follows.

1. complete an initial questionairre that provides demographic data
2. for each of two search methods assigned, do the following
   (a) go through a tutorial about that method
   (b) do several practice searches
   (c) for each of 2-3 questions provided, endeavor to carry out as many comprehensive searches as seem appropriate to find all relevant books listed in the catalog
3. complete a final questionairre regarding the overall system performance

Each search involves filling in blanks on the terminal screen that allow specification of a query of the proper form, examining the list of titles that are identified, examining detail records on titles of interest, and indicating which titles seem relevant to the question.

The proposed experiment is considered risk free to the participants, and only requires their typing into a computer. They will provide some information regarding their background so that we can determine if some search methods are better for some classes of users. They also will be making judgments regarding how to describe an information need (question) using this software system, and regarding whether retrieved book descriptions are for relevant books — so we can determine how well various retrieval methods perform regarding quickly finding relevant items.

Participants will all be University students, and we will endeavor to obtain a representative sample of the student community. This will give a relatively homogeneous group, but with enough diversity that we may identify trends regarding educational level, major, or other characteristics.

Participants will each have an ID number that we assign, but the log that associates names with ID numbers will be destroyed after data collection for the experiment is complete. Subsequent

analysis will be done for the group as a whole, and student-provided information will not be distinguishable for individual identification. Names will also be recorded, along with signatures, for all who receive payment, but ID numbers will not be used with that log.

After completion of the study and subsequent analysis, a technical report will be prepared describing this experiment, and copies will be available to the public (including students who have been participants). A paper will be submitted for publication to share our findings with the scientific community.

D Consent Form — REVTOLC (Retrieval Experiment — Virginia Tech OnLine Catalog)

Edward A. Fox, Project Director, Department of Computer Science
Virginina Polytechnic Institute and State University, Blacksburg, VA 24061-0106

I,_____, consent to participate in research concerning advanced methods of information storage and retrieval and their application to online public access library catalogs. I am aware that the purpose of this research is to compare various retrieval methods and to determine how efficient, effective, and usable they are by students — so that in the future it may be possible to have better catalog systems.

I acknowledge that my participation in this experiment is completely voluntary and will take roughly 60-90 minutes. I do reserve the right to quit at any time without penalty or retribution. I do realize, however, that $5 payment will not be made to me unless and until I have completely gone through all steps involved in the experiment.

I understand that the complete experiment for a given participant involves reading a brief explanation booklet, and then interacting through a terminal with a computer system. That interaction will include an initial questionairre, a final questionairre, and shorter intermediate questionairres after searches. I understand that I will work with two different search methods so that comparisons will be possible, and that for each of the two I will go through a brief tutorial, some practice searches, and then no more than 3 searches for various questions that will be assigned to me.

I am aware that my keystrokes and responses will be logged by the computer for future analysis aimed at comparing the retrieval methods and cross tabulating performance with characteristics of searchers. I have been assured that any personal information I provide as a part of this experiment will be kept confidential. I also understand that this information will be compiled for group analysis and used for statistical purposes only, and that the results of my participation will in no wise be attributed to me as an individual.

I am convinced that participating in this study poses no risks to my well-being. I expect that my participation will contribute to research in information retrieval and may lead to more effective catalog systems.

_____

Signature

_____

Date

E Explanation

REVTOLC (Retrieval Experiment with Virginia Tech OnLine Catalog)

Edward A. Fox, Project Director
Department of Computer Science
Virginina Polytechnic Institute and State University
Blacksburg, VA 24061-0106

E.1 FORMS

Thank you for volunteering to help with our experiment! If you are willing, before proceeding any further please sign the consent form you have been given, and hand it to the person running the experiment. Later, when you have completed the tasks assigned to you in the pages below, inform the person running the experiment of that fact, and you will be asked to sign another form indicating that you have been paid $5 for your time and cooperation.

Speaking of forms, PLEASE READ THE REST OF THIS EXPLANATION BOOKLET IN ITS ENTIRETY BEFORE TOUCHING YOUR ASSIGNED COMPUTER TERMINAL!!

E.2 PURPOSE OF THE EXPERIMENT

As was explained on the consent form, we are trying to compare how well various information retrieval methods help people searching for relevant entries in an online library catalog. You will work with two methods so we can see which one you like better, and then can correlate that with opinions of others participating in the experiment. For each method you will work through a tutorial and do a practice search, prior to searching for entries that satisfy each of two questions asked by other students who visited Newman library during this past year. How well you do and what comments you make about each search and each method will help us reach conclusions in our study, hopefully leading to better methods for searching through online library catalogs.

E.3 BACKGROUND

You may have looked for books in libraries using an old style card catalog. In recent years, the information in many of those catalogs has been put into computers so that people can more quickly find relevant books, and can easily make comprehensive studies to find as many items as possible on a subject (e.g., when planning research to be sure no one else has found the solution to a problem before). For example, at VPI&SU there is VTLS, a type of online catalog system.

Our system is somewhat like VTLS. At the end of 1986 we were given data by people in Newman Library so we could experiment with different methods of searching. That data, roughly 500,000 entries from the VPI&SU library that were cataloged by 1986, is accessible through REVTOLC. Since we are only studying about searching, and since the data we received is old, we only consider descriptions about the books and other entries – i.e., we ignore circulation records.

E.4 SHORT TUTORIAL ON SEARCHING

Finding entries using an online catalog usually involves several actions:

1. think a little about what you are looking for, namely your question

2. tell the computer something about that, in the form of a "query"

3. tell the computer to search using that query

4. examine the results

5. if you want more and believe more can be found, repeat steps 2-4

In this experiment, you will be given questions asked by other students, recorded on forms we collected in 1989 in Newman Library. Use whatever knowledge you have to help you understand what is meant. A good way to think about this is as follows: A friend of yours wants to find everything available in the library to do a comprehensive report, and writes down for you the question he is investigating. You want your friend to get a good grade, and so you must be as persistent and clever as you can, to tell the computer what it needs, so it can do searches. You must continue doing so until you have found everything you can relevant to the question. Since this could go on forever if you are really thorough, don't spend more than about 15 minutes on any question, and try to find entries "relevant" to the question, not just things that might be interesting.

Some librarians receive special training in searching, but you can do fairly well if you remember the following important points:

1. you must give the computer the right words or names

2. once you find something relevant, you can use it to get more words and/or names

3. you probably need to do at least 2-3 different searches for each question

Getting the right words or names takes thought. The problem is easier if you remember that librarians divide up descriptions of books into separate categories or "fields," for author, title, subject, call number, etc. If you know the name of an author, tell the computer to look for that in the author field. If you know words in the title, tell the computer to look for that in the title field. If you know the first part of the call number, that should be looked for in the call number field. If you know what Library of Congress subject heading would be assigned, that should be looked for in the subject field.

If you just know a little about the subject of the book desired, things are not so easy. Probably the safest strategy is to ask the computer to search for those words, looking in both the title and subject heading fields.

Remember point (2) above also! If you find some book that is relevant, you can use the author, subject heading, title, etc. in a follow-up search. It is likely that the author might have written something else similar. Other books classified with the same subject heading are likely to be related. Also, works with similar titles are probably useful. Call numbers can also be helpful, though it may not be very helpful to put in a call number like "QA76" which is used to describe thousands of engineering books!

One last point has to do with word forms. In the title field, this computer system will treat words with or without plural endings as being the same – so if you ask for "book" in the title and the author had "books" you will still get a "match." Furthermore, in the subject field, since people often don't know exactly what is counted as a subject heading by the Library of Congress, matches are generally considered as proper regardless of the endings. Thus "learning" would match "learns" – you may find some things you don't want because of this, but at least it is less likely that you will miss something you want!

E.5 USING THE COMPUTER

You will be using a terminal that will act like a VT100 terminal, since those are easy to find. This terminal is connected to a computer that has the catalog data stored so you can search. The computer is not very big, so be patient if it is a little slow at times.

The computer will tell you or show you things, and you should read whatever it says carefully. In some cases you will be asked to reply with a number, in which case you enter the number and then hit the RETURN key. In most other cases, you need only hit one letter or another key to get the computer to continue. What keys you can hit is listed in the upper right part of the screen, which changes as you proceed, since certain things only make sense at certain times.

The computer lets you use the backspace key to make corrections. When you are entering your query into the computer, the special key PF2 allows you to switch between overwriting what you typed before, or inserting newly typed characters. Also note that when you have finished with typing in a query, the special key PF1 is used to tell the computer to search for books.

If you are confused or unclear about what to do, ask one of the people running the experiment for help!

E.6 EXAMINING RESULTS

After the computer has searched based on one of your new queries, it will return with a list of entries, numbered 1-10, showing the title of each. If any title looks like it might possibly be relevant, please follow the instructions to examine the detailed description of that entry, and then indicate if you think it is indeed relevant. Remember that you should act as if your friend's report grade may depend on your persistence!

If you search again for the same question, after each such search the computer will find 10 more items, calling those 1-10, moving the others found earlier down in rank. You can still look over things found earlier, but new items are right at the top, easy to find.

E.7 FILLING IN QUESTIONNAIRES

There are a number of questionnaires that will be given online. Please take your time on these, and answer honestly, picking the most accurate reply at each point. At the very end of the experiment, in the final questionnaire, there will be a chance for you to write in more detailed comments of your own, so then you will be able to go beyond multiple choice replies!

E.8 YOUR MAIN TASK – WORKING WITH 2 OF 4 RETRIEVAL METHODS

You are assigned to work with two retrieval methods. Here is a short explanation of the four methods being considered in REVTOLC. No one knows how these compare when used for searching online catalogs. See more details about the two you will work with when you go through the online tutorials.

Boolean: Many retrieval systems allow you to precisely characterize what you are looking for by indicating combinations and alternatives of words or names. An example of a Boolean query is:

(Author: Gerard AND Salton) OR (Title: information AND retrieval)

which will find books having an author with the two names "Gerard" and "Salton", or books that have "information" and "retrieval" in the title. No other books would be found. You would miss a

book by "G. Salton" and also would miss a book where the title was "Searching for Information". Thus, you will need to carefully think about the wording and combinations to form a Boolean query. Note finally that a Boolean query may not match any books, may match less than or equal to 10 (in which case you will be shown only those), or may match more than 10 (in which case you only will be shown 10 that the computer arbitrarily selects from among the matches).

You can manually alter your query after a search by adding or deleting words in the various fields on the screen. That way you can try to be narrower in your query formulation if you found lots of items, or give a broader query if you did not find much.

Note that the tutorial gives special instructions regarding filling in your Boolean query, and that the special keys PF3 and PF4 can be used to make that process easier. By hitting those keys repetedly you can make the computer toggle or alternate between using AND and OR to combine what is specified on different lines and what is specified in different fields.

Pnorm: This method is like Boolean. You should put in a query in much the same way. But the computer acts differently. It considers all books that have at least one of the words or names you have given in the query, and then ranks them based on how well it thinks they match what you asked for. Thus, it is more forgiving and can find relevant books that do not exactly match your query. On the other hand, it might find things you don't want. It uses special mathematical computations to rank what it finds, so the ones closer to the top of the list are likely to be the best.

Vector: Vector retrieval is very different from Boolean. The idea is to put in all words or names that you are pretty sure are in the books you want, and to leave the rest to the computer. It uses mathematical methods to treat your query as a list or vector of terms, and employs pattern matching techniques to find books that seems to have the best fit to your query. You don't need to worry about ANDs and ORs. The computer will try to rank all items that have any of the words you give, putting those with the best and most matches nearer the top of the list.

To make another search for a given query, manual feedback lets you add or delete words to the query you used before, and try again.

Vector with automatic feedback: This is very much like vector retrieval. The only difference is that there is another option besides manual feedback, which is available when you want to make another search. It is called automatic feedback.

It can be used if you marked one or more books with "y" to indicate relevance to the question. In automatic feedback, the computer looks at the descriptions of all books you marked as relevant to the question, and adds words and names from them to your query. It uses special mathematical computations to try to find more books like the resulting query (that has been enhanced with these new words and names). Thus, automatic feedback is a way for the computer to relieve you of the necessity to make a new query.

You have been assigned to use [filled in with METHOD1] as your first search method, and to use [filled in with METHOD2] as your second search method.

E.9 FINAL COMMENTS, SUMMARY

1. Ask for help from a person running the experiment if you are confused or want assistance regarding how the computer is behaving or how to get it to help you better.

2. On each of the 4 assigned questions, don't plan on spending more than 15 minutes per question unless you really want to.

3. Note that you will now do the following on the computer:

(a) When asked, type in your group number: [filled in with GROUP]

(b) When asked, type in your ID number: [filled in with ID]

(c) Read more about the experiment and about using the computer

(d) Answer questions for the initial questionnaire

(e) Learn about your search method 1, through the online tutorial. Your method 1 is: [filled in with METHOD1]

(f) Do at least one practice search to get used to method 1. For that, you should look for books by Gerard Salton or other books about "information retrieval". If you want more practice, go ahead only if you are willing to spend more than the 60-90 minutes agreed upon. In that case make up a question of interest to you, and search.

(g) When you finish each search and go on to another search, reply to the questions in the search questionnaire.

(h) Still using method 1, do two new searches, and fill in the questionnaire that follows each search, for the following two questions:

  i. Question: [filled in with question 1 for this person]

  ii. Question: [filled in with question 2 for this person]

(i) When you are done searching with method 1, fill in the questionnaire about that method, and proceed to method 2.

(j) Repeat steps e-h for method 2, which for you is: [filled in with METHOD2], searching for the following two questions:

  i. Question: [filled in with question 3 for this person]

  ii. Question: [filled in with question 4 for this person]

(k) When you are done searching with method 2, fill in the questionnaire about that method, and then proceed to fill in the final questionnaire.

(l) Go to the person running the experiment, and when you show that you are done, sign the receipt sheet and receive your payment.

Now, go ahead and work with the computer, following instructions given above. Thank you for helping with our experiment!

F Questions

1. How many computer courses have you taken before college or in college
2. How often do you use a computer
3. Do you like to use a computer
4. How many times have you used VTLS
5. How many times have you used another retrieval system
6. What is your typing skill
7. Your gender is

8. You would describe yourself as
9. What is your overall grade point average
10. What is your academic level
11. The category that best describes your academic area is
12. The question you are trying to find documents to satisfy is
13. The time delay when the computer is finding documents is
14. This computer search has helped you to find
15. Among the citations provided by this search, what percentage seemed relevant to the question you entered
16. Which of the following best describes your search
17. In general, using this method to search for library material takes
18. The results of the searches were what I expected
19. To what degree do you think this method is an effective aid to casual searching for a few library references?
20. To what degree do you think this method is an effective aid to comprehensive searching for as complete a set as possible of relevant library references
21. Being able to manually edit your earlier query is helpful
22. The display of the details of documents retrieved is clear and adequate
23. Was the information and assistance given to you by the method easy to understand and adequate
24. It is easy to learn to use this method
25. In general, this method is easy to use
26. Searching with this method is easier than searching with VTLS
27. In general the method did a good job helping you find useful items
28. There is nothing else I want to say or comment about the method or this study
29. The editor for entering your question is easy to use
30. The feedback facility helped you find additional useful items
31. The computer has tried to rank output so best results are first – this ranking of output helps quickly identify relevant documents